Sample-Path Optimization of Buffer Allocations in a Tandem Queue - Part I

Gürkan, G.; Ozge, A.Y.

Publication date:
1996

Submitted for publication. October 15, 1996.

Sample-Path Optimization of Buffer Allocations in a Tandem Queue - Part I: Theoretical Issues

Gül Gürkan [1]
CentER for Economic Research
Tilburg University
P.O. Box 90153
5000 LE Tilburg, The Netherlands
email: ggurkan@kub.nl

A. Yonca Özge
Department of Industrial Engineering
University of Wisconsin-Madison
1513 University Avenue
Madison, WI 53706-1572, USA
email: ozge@cs.wisc.edu

Abstract: This is the first of two papers dealing with the optimal buffer allocation problem in tandem manufacturing lines with unreliable machines. We address the theoretical issues that arise when using sample-path optimization, a simulation-based optimization method, to solve this problem. Sample-path optimization is a recent method to optimize performance functions of stochastic systems. By exploiting the fact that the performance function we want to optimize is the almost sure limit of a sequence of random functions, it overcomes some of the difficulties from which variants of stochastic approximation methods suffer.

We provide a mathematical framework that makes use of a function space construction to model the dependence of throughput on buffer capacities and maximum flow rates of machines. Using this framework we prove various structural properties of throughput and show how these properties, along with a niceness condition on the steady-state, can be used to prove that the sample-path optimization method converges almost surely when applied to the buffer allocation problem.

Among the properties established, monotonicity in buffer capacities and in machine flow rates are especially important. Although monotonicity results of this nature have appeared in the literature for *discrete* tandem lines, as far as we are aware the kind of analysis we present here has not yet been done for *continuous* tandem lines.

Key Words: Stochastic optimization, buffer allocation, tandem manufacturing lines, steady-state throughput, monotonicity, sample-path optimization.

[1] This work has been done when Gül Gürkan was at the Department of Industrial Engineering, University of Wisconsin-Madison.

The research reported here was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant number F49620-95-1-0222. The U. S. Government has certain rights in this material, and is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the sponsoring agency or the U. S. Government.

1 Introduction

Research reported in the present and the accompanying paper aims to enhance the set of available tools for analyzing and optimizing tandem lines with unreliable machines. In this paper we provide a novel mathematical framework to model the dynamics of such systems, propose a new solution methodology for the buffer allocation problem, and discuss the technical details of our solution. The accompanying paper, Gürkan (1996b), will discuss operational issues that arise during the implementation.

A tandem queue consists of a number of servers in series. There may be buffers of finite sizes between the servers. Jobs start at the first server, pass through each server in sequence, and finally leave the system after being served by the last server. These queues have been widely used to model a single line of multistage automated assembly lines or virtual paths in communication networks; see for example Ho *et al.* (1983), Gershwin (1987), Buzacott and Shanthikumar (1992), and Yamashita and Önvural (1994).

We focus on a particular tandem queue where service rates are deterministic. The servers are subject to random breakdowns: these failures are operational (i.e. a server may only breakdown during service); operating quantities between failures as well as repair times are random with arbitrary distributions. It is common to use this type of tandem queues to model tandem production lines in which machines are the servers, see Buzacott and Shanthikumar (1992) and Yamashita and Önvural (1994) and references therein. In a tandem production line, the material processed may be discrete entities (e.g. assemblies in an automobile factory) in which case we call it a discrete tandem (DT) line or it may be continuous (e.g. chemical production) in which case we refer to a continuous tandem (CT) line.

Possible decision variables in tandem production lines include buffer capacities, cycle times of machines, and failure and repair rates of machines. Recently, there has been progress towards the optimization of steady-state throughput, the amount of production per unit time by the last machine in steady-state, with respect to machine *cycle times*. Plambeck, Fu, Robinson, and Suri (1996) used the sample-path optimization method to optimize lines with up to 50 machines under various linear equality and inequality constraints on the cycle times. The aim of the current paper and its companion Gürkan (1996b) is to take this one step further and to optimize the steady-state throughput with respect to *buffer capacities*. Under certain conditions, the existence of a steady-state in tandem queues is guaranteed by regeneration theorems. We do not go into any detail about such conditions; we refer the reader to, for example Loynes (1962), Nummelin (1981), and Gershwin and Schick (1983).

The buffer allocation problem is still an open question in the study of tandem production lines. Analytical results based on Markov chain representations of the model are available only for 2- and 3-machine DT lines in Gershwin and Schick (1983), and for 2-machine CT lines in Gershwin and Schick (1980). To find optimal buffer al-

locations in DT lines, a heuristic method based on a Markov chain representation was used in Hillier and So (1991); since the number of states of the Markov chain grows very rapidly with increasing number of machines and buffer capacities, only lines with up to five machines could be considered. The intractability of analytical models for long production lines makes simulation an attractive approach to study these lines. A method to estimate the gradient of line throughput with respect to buffer capacities in DT lines was introduced in Ho *et al.* (1979) and these gradient estimates were then used in a heuristic "hill climbing" algorithm to find optimal buffer allocations. This was also the first paper in which the technique of perturbation analysis was used to compute gradient estimates in discrete-event dynamic systems. As for CT lines, an algorithm based on generalized Benders' decomposition was developed to optimize steady-state throughput and in-process inventory with respect to buffer capacities in Caramanis (1987). To compute the necessary gradients, the approach of Ho *et al.* (1979) was adopted. There was no justification for using a deterministic optimization technique with noisy function and gradient values to solve a stochastic optimization problem.

One approach to model and analyze DT lines is to approximate them by CT lines. The continuous production case can be visualized as the limit of the discrete production case as the piece size approaches zero while the production rate remains constant; see Fu (1996). For a translation of various input parameters and performance measures between CT and DT lines, see Suri and Fu (1994). A major reason for using CT lines instead of DT lines is the considerable increase in computational efficiency. Extensive numerical results on the substantial time advantage of CT simulations over DT simulations are reported in Suri and Fu (1994). Using CT lines is beneficial from an optimization point of view as well: techniques for continuous parameter optimization are much more advanced than those for discrete parameter optimization. Furthermore, when dealing with continuous parameters there is the possibility of obtaining gradient estimates.

For these reasons, to optimize the steady-state throughput with respect to buffer capacities, we will use CT line approximations for DT lines. Extensive numerical results on both DT and CT lines in Suri and Fu (1994) indicate that approximation of DT lines via CT lines is quite accurate. For example, for fairly small lines (up to six machines), the throughput values obtained from CT line approximations in Suri and Fu (1994) were very close to the throughput of the original DT line (relative errors ranging from 0.0% to -2.3%); in an extensive study of 192 15-machine lines, in 90% of the cases the difference between the DT line throughput and the equivalent CT line throughput was less than 4%. Since CT line simulations are substantially faster than DT line simulations and the approximations are quite accurate, we believe optimizing CT lines is an important step in enhancing the set of tools available for optimizing DT lines.

The buffer allocation problem in tandem production lines is one instance of a

generic simulation optimization problem: given that one can obtain a function and a gradient value at a parameter setting, locate an optimizer of the performance function. When faced with this problem, people often used some form of the stochastic approximation method; see Robbins and Monro (1951), Kiefer and Wolfowitz (1952), or the single-run optimization variant Meketon (1983, 1987). These methods are known to have a number of drawbacks. First, their empirical performance is very sensitive to the choice of a predetermined step size. Fu and Healy (1992) and L'Ecuyer *et al.* (1994) contain a number of examples which demonstrate this sensitivity. Second, since they are mainly first-order gradient methods, they are often thought to experience more difficulties on large problems than on small problems. Third, in case of constrained optimization, these methods handle inequality constraints –even linear inequalities– via projection onto the feasible set. This can retard the performance of an algorithm immensely, as is illustrated by an example in Appendix 6 of Plambeck, Fu, Robinson, and Suri (1996). In that example, such a method requires nearly 10^{43} steps to find the minimizer (the origin) of a linear function on the nonnegative orthant $\mathbf{R}^2_+$. Notice that this difficulty does not arise in case of linear equality constraints since one can reduce this to an unconstrained problem by appropriate affine transformations. Finally, if the function being optimized is non-differentiable, then the stochastic approximation method becomes a variant of subgradient optimization; see Correa and Lemaréchal (1993) for example. That method is known to be very slow and it also suffers from other drawbacks such as the lack of a good stopping criterion and the difficulty in enforcing feasibility as mentioned above.

Recently a new method called *sample-path optimization* that overcomes some of these difficulties was proposed in Plambeck, Fu, Robinson, and Suri (1996) and analyzed in Robinson (1996). The method exploits the fact that the performance function we want to optimize is the almost-sure limit of a sequence of approximating functions (outputs of simulations of runs of increasing lengths, all using the same random number streams). That is, if we go out far enough along the sample path we get a good estimate of the limit function. Being a deterministic function, this resulting estimate can then be optimized using deterministic optimization techniques. One of the most powerful features of sample-path optimization is the availability of superlinearly convergent (fast) deterministic optimization methods that can handle constraints explicitly and that do not suffer from increases in the problem dimension. Using these methods we can often optimize the approximating function to high accuracy in relatively few function and gradient evaluations. This is particularly important when function and gradient evaluations are expensive. The method can be used even when the performance function or the sample functions are non-differentiable (the convexity of the functions is required in this case), this time using methods of non-smooth convex minimization, such as bundle algorithms, in the optimization scheme. In addition, the method separates optimization from the compu-

tation of function and gradient values. This modularity turns out to be quite helpful when the system simulated is large and complex and/or the optimization routine is sophisticated.

Since the optimization problem we are facing is a difficult one with possibly several variables and constraints, sample-path optimization is evidently a promising stochastic optimization technique to solve this problem. In §2.2 we describe the sample-path optimization method in more detail. Appendix A contains the convergence results of the method which are relevant to the work presented in this paper. For proofs and detailed analysis of those results, we refer the interested reader to Robinson (1996). For successful applications of the method on systems of considerable sizes, see Plambeck, Fu, Robinson, and Suri (1996). A comprehensive summary of the properties of the method is given in Gürkan, Özge, and Robinson (1994) which also reports the performance of the method on a small closed queueing network. An alternative set of conditions to those developed in Robinson (1996) for proving the convergence of the method are provided in Gürkan, Özge, and Robinson (1996). This new set of conditions substantially broadens the class of problems to which the method is applicable; in particular it enables the solution of stochastic variational inequalities using the sample-path technique. A brief survey of related techniques and ideas similar to sample-path optimization that have appeared in the literature can be found in Robinson (1996).

The remainder of this paper is divided into four main sections. At the end of the paper there are four (or three) appendices containing additional technical detail. Of the main sections, Section 2 contains the description of the problem and the solution methodology we propose. In §2.1, we describe the characteristics of the tandem line under study. In §2.2, we discuss the basic ideas behind the sample-path optimization method. In §2.3, we mention the advantages of using sample-path optimization to find optimal buffer allocations and discuss some of the issues associated with this approach. Section 3 is devoted to addressing the theoretical issues. In §3.1, we provide a mathematical framework to model the dynamics of the tandem line and develop the necessary machinery for the technical analysis. In §3.2, we prove some properties of throughput, namely monotonicity, upper semicontinuity, and properness. Among those properties, monotonicity in buffer capacities and in machine flow rates deserve special attention. Although monotonicity results of this nature for DT lines have appeared in the literature, (best to our knowledge) such results were not available for CT lines. After establishing these properties, we then show how they can be used to prove the convergence of the sample-path optimization method when applied to the buffer allocation problem. Finally, in Section 4 we summarize the work presented, briefly mention some of the practical issues that arise when implementing the method. For a complete discussion of these operational issues we refer the reader to Gürkan (1996b).

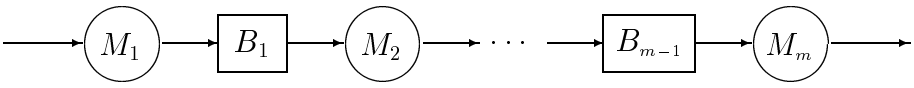

Figure 1: The tandem production line

2 Description of the problem and our approach

2.1 The buffer allocation problem

A tandem line consists of m machines in series connected by $m-1$ buffers of possibly finite sizes. The product enters from one end of the line, goes to each machine in sequence, finally emerges from the other end as a final product. The time it takes a machine to process one unit of product is called the cycle time. Notice that in a CT line the natural description for the processing rate of a machine is the flow rate which is the reciprocal of cycle time.

Tandem lines are a class of production lines which are commonly used for mass production of various products. The study of such lines may be required in order to design a new line or to improve an existing line. In either case, we are faced with an optimization problem in a complex stochastic system: to optimize the performance of the line under various financial and/or non-financial constraints. Possible decision variables include buffer capacities, cycle times, and failure and repair rates of machines. There are many trade-offs arising from the complex dynamics of the system which make analytical study of these lines very difficult. People who study these lines usually focus on two common performance measures: line throughput, the amount of production per unit time, and in-process inventory. In this paper, we focus on the line throughput; since the in-process inventory is bounded by the total buffer capacity, and the cost associated with it can be incorporated into the overall buffer cost.

The particular tandem line we study has the following additional features:

- There is infinite supply to the first machine and infinite demand from the last machine.
- There is no transfer delay from machines to buffers, within buffers, or from buffers to machines.
- A machine can fail only when it is operational. Operating quantity to failure for each machine is a random variable.

- The repair time for each machine is a random variable.
- Each machine has a deterministic maximum flow rate, so each machine can work at a rate anywhere between zero and this maximum.

A few words about the dynamics of the line are in order. Consider a machine M_i. Since it is unreliable, it will sometimes fail. As soon as it is repaired it will continue to produce until another failure occurs. In addition to its own failures, sometimes M_i may have to reduce its production rate or even completely stop because of the interactions with other machines. For example,

a) If the buffer B_i is full, M_i cannot produce at a rate larger than the current rate of M_{i+1}. In such a case M_i is said to be *blocked*.

b) Similarly, if the buffer B_{i-1} is empty, M_i cannot produce at a rate larger than the current rate of M_{i-1}. In such a case M_i is said to be *starved*.

These characteristics result in complex dynamics for the system and have made it impossible (to date) to use analytical methods to optimize the performance measures such as steady-state throughput or in-process inventory. In this work we focus on optimizing the steady-state throughput with respect to buffer capacities under various constraints. As a result of the interactions between the machines, one would like to increase the buffer capacities to make the machines more independent of each other to increase the throughput. However, due to financial and spatial limitations, increasing the buffer capacities may not be feasible.

For the tandem lines described above, analytical expressions for steady-state throughput of 2-machine CT lines and of 2- and 3- machine DT lines are available; see Gerschwin and Schick (1980) and the references in Suri and Fu (1994). However, lack of analytical results for longer lines makes using simulation attractive to analyze and optimize these lines. Sample-path optimization is a powerful simulation-based method that can be used in the solution of this problem.

2.2 Sample-path optimization method

In this section we describe the basic ideas behind a simulation-based method, sample-path optimization, for optimizing performance functions in certain stochastic systems. We do not go into any technical detail and refer the interested reader to Robinson (1996). However, since in §3.2 we make use of the main convergence result of that work, we provide it in Appendix A.

Many problems in simulation optimization can be modeled by an extended-real-valued stochastic process $\{L_n(\theta) \mid n = 1, 2, \dots\}$. The L_n take values that may be real numbers or $\pm\infty$, whereas the parameter θ takes values in $\mathbf{R}^k$. Using extended-real-valued random variables is very convenient for modeling constraints, since one can always set $L_n(\theta) = +\infty$ for those θ that do not satisfy the constraints. For each

$n \geq 1$ and each $\theta \in \mathbf{R}^k$, $L_n(\theta)$ are random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$.

The method assumes the existence of a limit function L_∞ such that the L_n almost surely converge pointwise to L_∞ as $n \to \infty$. For the systems we are concerned with, such existence and convergence can often be inferred from regeneration theorems and/or the strong law of large numbers. In the following we refer to $L_n(\theta)$ as the sample function and we write $L_n(\omega, \theta)$ when we want to emphasize the dependence of $L_n(\theta)$ on the sample point ω.

Let us demonstrate this setup with a simple example. Suppose that we are analyzing an $M/M/1$ queue and we are interested in the steady-state system time of a customer, denoted by L_∞. Let L_n be the average of the system times of n customers, i.e. L_n is the output of a simulation of run length n (n service completions in this case). ¿From the regeneration theorems we know that under certain conditions on the parameters of the system L_∞ exists and the L_n converge pointwise to L_∞ along almost every sample path.

We are interested in finding the infimum and, if it exists, a minimizer of L_∞. In general we can only observe L_n for finite n. Therefore we approximate minimizers of L_∞ using such information about L_n. The method is simple: fix a large n and $\omega \in \Omega$, compute a minimizer $\theta_n^*(\omega)$ of $L_n(\omega, .)$, and take $\theta_n^*(\omega)$ as an approximate minimizer of $L_\infty(\omega, .)$. Note that minimizers of $L_\infty(\omega, \cdot)$ may generally depend on the sample point ω. However, in many practical problems for which one would anticipate using this technique L_∞ is a deterministic function, for example a steady-state performance function or an expected value, i.e. it is independent of ω.

As shown in Robinson (1996), the conceptual method of sample-path optimization converges with probability one under three hypotheses: the approximating functions $L_n(\omega, \cdot)$ are lower semicontinuous and proper; they epiconverge to the limit function $L_\infty(\omega, \cdot)$; and the limit function $L_\infty(\omega, \cdot)$ almost surely has a nonempty, compact set of minimizers. For a precise statement of this result, see Theorem 8 and Proposition 2 in Appendix A.

Notice that once we fix n and a sample point ω, $L_n(\omega, \theta)$ becomes a deterministic function of θ. With this observation, very powerful methods of constrained and unconstrained deterministic optimization are available to use on L_n. In the smooth case we can apply superlinearly convergent methods like the BFGS algoritm (or a variant of it in case of constraints) to minimize L_n to high accuracy in few function and gradient evaluations. For more information on these algorithms see Fletcher (1987) and Gill *et al.* (1981) and for the software available see Moré and Wright (1993). Use of superlinearly convergent methods enables us to be confident about the location and the accuracy of the minimizer of L_n; i.e. we can differentiate between the errors due to the approximation of L_∞ by L_n and those due to the inaccurate computation of a minimizer of L_n. With slower algorithms like stochastic approximation this is difficult, if not impossible. If the sample functions and/or the performance function

we want to minimize are nondifferentiable and convex, then we can use the Bundle-Trust method; see Schramm and Zowe (1990) and Kiwiel (1990). We emphasize that in both the smooth and the non-smooth case, the deterministic solution methods available can handle constraints explicitly and without any difficulty.

Another useful feature of this approach is its modularity; the computation of function and gradient values is separated from the optimization. This enables the use of already existing simulation codes (if they also provide gradient values or can be modified to do so) together with optimization codes that call external subroutines for function and gradient evaluations. If the system simulated is large and complex, and the optimization code is sophisticated, then the advantage of modularity becomes more substantial.

2.3 The solution methodology

Recall that our objective is to optimize the steady-state throughput with respect to buffer capacities under various constraints, using a simulation-based optimization method. To solve this problem with the existing technology, namely with variants of the stochastic approximation method, for CT lines of even moderate sizes is inefficient and difficult (if not impossible). This is due to several drawbacks that stochastic approximation methods suffer from, including lack of a good stopping criterion, difficulty in enforcing feasibility, and slow empirical convergence rate as discussed in Section 1. As mentioned above, sample-path optimization is a recent alternative that is suitable for optimizing performance measures of complex systems CT lines. Furthermore, the method has been tested numerically on a number of applications and the computational experience to date has been very promising. In all cases, computational results suggest that even a fairly small computational effort may produce a solution that is accurate enough for practical purposes: substantial increases in computation time resulted in fairly small changes in the optimal solution.

Using sample-path optimization to solve the buffer allocation problem has two apparent advantages; the effect of modularity will be quite substantial due to the size and the complex dynamics of the system and the availability of superlinearly convergent deterministic optimization algorithms will enable us to locate the optimizer to high accuracy in relatively few function and gradient evaluations even in the presence of numerous and/or complicated constraints.

Clearly, this idea raises a number of questions:

(i) What are the theoretical issues that arise when we attempt to apply sample-path optimization to the buffer allocation problem? Is the problem well structured enough?

(ii) What are the operational issues we have to deal with, if this method is used to solve a real-world problem?

(iii) Is it computationally feasible to use this method and if so how well does it perform in this particular application?

It turns out that answering the first set of questions is closely linked to establishing certain properties of the function to be optimized. In an actual problem, to optimize the performance of the system, one would minimize a combination of reciprocal of throughput and a cost function. The cost function usually captures some information regarding the financial limitations and space availability about buffer capacities. Provided that the added cost function has a reasonable functional form (e.g. continuity), the properties of the function we want to minimize will follow from the properties of throughput. In Section 3, we establish certain properties of throughput and discuss their implications for the method's convergence.

In summary, the method we propose consists of optimizing the deterministic function obtained by fixing a sample path. In §3.2, we show that under a regularity condition on the steady-state, the optimizer computed using such a scheme converges almost surely to the correct optimizer as we go far enough on the sample-path. This makes us confident about using sample-path optimization to find optimal buffer allocations. At the next stage of this work, using CT line simulations enables us to compute certain directional derivatives using infinitesimal perturbation analysis (IPA) from a single realization of the sample path. We then locate a minimizer of the resulting function using the most powerful deterministic optimization techniques available to us. A detailed discussion of these implementation issues can be found in Gürkan (1996b).

3 Technical analysis

In this section we develop the machinery required to deal with the theoretical issues that arise when applying sample-path optimization method to find optimal buffer allocations. In §3.1 we provide a mathematical framework to model the dynamics of the tandem line and in §3.2 we use this framework to prove various properties of throughput and the convergence of the conceptual method when applied to the buffer allocation problem.

3.1 Mathematical framework

Let T be the prespecified amount of time we observe the line and $q_i(t)$ be the amount produced by M_i up to time t for $i = 1, \ldots, m$. Then the line throughput can be defined as

$$TP_T = q_m(T)/T.$$

We define

b_j = buffer capacity of B_j,

C_i = maximum flow rate of M_i,
W_p^i = operating quantity between the $(p-1)st$ and the pth failures at M_i,
R_p^i = repair time of M_i after the pth failure.

For each i, $\{W_p^i\}_{p=1}^n$ and $\{R_p^i\}_{p=1}^n$ are random variables with distributions that are concentrated on $(0,\infty)$.

For a fixed sample path, i.e. for fixed sequences $\{W_p^i, i=1,\dots,m, p\geq 1\}$ and $\{R_p^i, i=1,\dots,m, p\geq 1\}$, let f_{ij} be the quantity produced by the ith machine up to its jth failure. Then

$$f_{ij} = \sum_{p=1}^{j} W_p^i.$$

Fix T (simulation time), let $\mathcal{C}([0,T],\mathbf{R}^m)$ be the space of continuous functions from $[0,T]$ to $\mathbf{R}^m$ with the sup-norm topology. That is, for $g\in\mathcal{C}([0,T],\mathbf{R}^m)$,

$$\|g\| = \sup\{|g_i(x)| : i=1,\dots,m, x\in[0,T]\}.$$

We next construct a multifunction $F : \mathbf{R}^{m-1} \to \mathcal{C}([0,T],\mathbf{R}^m)$ as follows. For any $b=(b_1,\dots,b_{m-1})\in\mathbf{R}_+^{m-1}$, we define $F(b)$ to be the set of continuous functions $g:[0,T]\to\mathbf{R}^m$ satisfying the following requirements:

$g_1\geq g_2\geq\ldots\geq g_m\geq 0$,
g_i is non-decreasing for each $i=1,\dots,m$,
$g(0)=0$,
$|g_i(x)-g_i(y)|\leq C_i|x-y|$ for any $x,y\in[0,T]$ and $i=1,\dots,m$,
$g_i(x)-g_{i+1}(x)\leq b_i$ for any $x\in[0,T]$ and $i=1,\dots,m-1$.

For $b\notin\mathbf{R}_+^{m-1}$, we let $F(b)=\emptyset$. Hence $\operatorname{dom}F=\mathbf{R}_+^{m-1}$. The graph of F is defined as $gphF=\{(b,g): g\in F(b)\}$. One should think of the functions $g\in F(b)$ as possible ways of operating the CT line. If we interpret $g_i(t)$ as the amount produced by machine i up to time t, then functions in $F(b)$ obey the buffer capacity and maximum flow rate constraints:

(i) the amount produced by a machine cannot be less than the amount produced by the succeeding machine,

(ii) the amount produced by a machine does not decrease with time,

(iii) the line starts operating at time zero,

(iv) a machine cannot work at a rate higher than its maximum flow rate,

(v) the amount produced by a machine cannot exceed the amount produced by the succeeding machine plus the buffer capacity between them.

We define A to be the following subset of $F(\infty)$:

$$A=\{g\in F(\infty) : \lambda(\{t : g_i(t)=f_{ij}\})\geq R_j^i, \text{ for each } i=1,\dots,m \text{ and } j=1,2,\dots\},$$

where λ is the Lebesgue measure on $\mathbf{R}$. Again, if we think of functions in A as possible ways of operating a CT line with unlimited buffer capacities between machines, then

the condition $\lambda(\{t : g_i(t) = f_{ij}\}) \geq R_j^i$ means that under any possible operating strategy the amount of time machine i stays non-operational after its jth failure is at least equal to its jth repair time.

$F(b)$ models the buffer capacity and maximum flow rate constraints, whereas A models the failure and repair times of a CT line with unlimited buffer capacities between machines. So $F(b) \cap A$ can be thought of as the set of all possible ways of operating the CT line. Notice that q is in $F(b) \cap A$. Recall that among functions in $F(b) \cap A$, q gives the amount produced using the strategy under which each machine is operated at maximum possible rate whenever it is operational. The pseudo-code developed in Fu (1996) prescribes a way of constructing such a strategy during a simulation. His v_i is the effective flow rate of M_i, $i = 1, \ldots, m$; at any time the pseudo-code prescribes how to set each one to its maximum possible value in a well-defined, non-circular way.

Using this framework we can have the following three technical lemmas; their proofs are deferred to Appendix B.

Lemma 1 *The multifunction F has the following properties:*

a. $\operatorname{gph} F$ *is closed.*

b. $\operatorname{gph} F$ *is convex.*

c. F is compact-valued and $F(b) \subset F(\infty)$ for all $b \in \mathbf{R}^{m-1}$.

In the next lemma we use the terms Berge-usc and Hausdorff distance, which are defined in Appendix A. We denote the *interior* of a set S by $\operatorname{int} S$.

Lemma 2 *The multifunction F is Berge-usc in $\mathbf{R}^{m-1}$ and $b \mapsto F(b)$ is a continuous mapping from $\operatorname{int}(\mathbf{R}_+^{m-1})$ to compact subsets of $\mathcal{C}([0,T], \mathbf{R}^m)$ with the metric topology induced by the Hausdorff distance.*

Lemma 3 *A is closed in $F(\infty)$.*

Now let

$$Q_T(b) = \sup\{g_m(T) : g \in F(b) \cap A\}.$$

In the next theorem, we show that the supremum in the definition of $Q_T(b)$ is actually attained and it is equal to the amount produced by the last machine up to time T when each machine is operated at maximum possible rate whenever operational. The proof of the theorem is provided in Appendix C.

Theorem 1 *Suppose that the event times have no cluster point. Then for each finite time T, $Q_T(b) = q_m(T)$.*

Remark The assumption that the event times have no cluster point is realistic since in any computer simulation of finite length the distinct random numbers generated are separated by some $\sigma > 0$ determined by the specifications of the computer.

3.2 Properties of throughput and their implications on convergence

We now discuss some properties of TP_T and TP_∞, namely upper semicontinuity, monotonicity (non-decreasing), and properness. As one can see from Theorem 8 of Appendix A, these turn out to be crucial in proving the convergence of the sample-path optimization method when applied to the buffer allocation problem.

Below we use the term "non-decreasing" for a function $f : \mathbf{R}^m \to \mathbf{R}$, by which we mean that $f(x_1, \ldots, x_k) \geq f(y_1, \ldots, y_k)$ whenever $x_i \geq y_i$ for $i = 1, \ldots, k$.

Theorem 2 *For $T \in [0, \infty]$, TP_T is a non-decreasing function of b with probability one.*

Proof. Observe that for $b' \leq b$, $F(b') \subset F(b)$. Hence $Q_T(b') \leq Q_T(b)$ and TP_T is a non-decreasing function of b. ∎

The reader may compare this monotonicity result with Meester and Shanthikumar (1990). Their paper is concerned with monotonicity of throughput as a function of buffer capacities of a *discrete tandem queue* with exponential service times, whereas we are concerned with monotonicity of throughput of a *continuos tandem line* with unreliable machines and deterministic flow rates. Furthermore, we do not make any distributional assumptions for the failure and repair times. Aside from these differences, our proof technique is quite different from theirs. They use certain recursive equations to characterize the dynamics of the system, especially the number of departures from each server, and obtain the result by manipulating these equations inductively, whereas we provide a new function space representation to model the dynamics of the system and exploit this mathematical framework to obtain the result.

Meester and Shanthikumar (1990) and Anantharam and Tscoucas (1990) also show the concavity of sample throughput in buffer capacities. This result holds for the discrete analog of the system we are studying if failure and repair times are exponentially distributed, as shown in Gürkan (1996a); however it fails to hold for CT lines; see Figure 2 and the discussion following Theorem 4.

One can also define a multifunction $\tilde{F}(C)$ from $\mathbf{R}^m$ to $\mathcal{C}([0, T], \mathbf{R}^m)$ by the same four conditions that we used to define F, where the variable is C, the vector of maximum flow rates, and follow the lines of proof of Theorem 2 to prove the monotonicity of throughput in flow rates. We note that though it is not the subject of the work we report here, the mathematical framework provided in §3.1 may facilitate similar analysis for throughput as a function of flow rates.

Theorem 3 *For $T \in [0, \infty]$, TP_T is a non-decreasing function of C with probability one.*

Proof. Observe that if $C' \leq C$, then $\tilde{F}(C') \subset \tilde{F}(C')$. ∎

We should point out the difference between the monotonicity result of Theorem 3 and those of Shanthikumar and Yao (1989a); as in the previous result the difference is in the system studied and the proof technique employed. Theorem 3 is concerned with continuous tandem queues, whereas Shanthikumar and Yao study general discrete queueing networks for which the discrete tandem queue is a special case and use recursive equations to establish the monotonicity of throughput in the job service times. In addition to monotonicity, Shanthikumar and Yao (1989b) show that the reciprocal of throughput is a convex function of parameters of the external interarrival times and the machine service times, provided that these times themselves are convex functions of those parameters. This convexity result is later extended to discrete tandem queues with unreliable machines in Fu (1996). In addition, the convexity of reciprocal of throughput in maximum flow rates of machines in CT lines is proven in Fu (1996). We pointed out to B.-R. Fu that by using the recursive equations for departure time process developed in Fu (1996), he can also show the monotonicity of throughput in maximum flow rates of machines; this would be an alternative way of proving Theorem 3.

Remark A generalized semi-Markov process (GSMP) representation is constructed in Suri and Fu (1994) to model CT lines. In Gürkan (1996b), it is shown that this GSMP is *not* non-interruptive (in the sense of Schassberger (1976)). Unfortunately, violation of the non-interruption condition rules out the applicability of the results, developed in Glasserman and Yao (1992a, 1992b), for checking the first and second order properties of stochastic systems that can be modeled as non-interruptive GSMP's.

The next result deals with the upper semicontinuity of sample throughput. This is important for two reasons: convergence analysis of the sample-path optimization method for our problem requires upper semicontinuity of sample throughput, and lack of upper semicontinuity in a function to be maximized may cause great difficulties when doing practical optimization.

Theorem 4 *For $T \in [0, \infty)$, TP_T is an upper semicontinuous function of b with probability one.*

Proof. Let $T \in [0, \infty)$. We will show that $q_m(T)$ is an upper semicontinuous function of b and the result will follow since $TP_T(b) = q_m(T)/T$. By Theorem 1 it is enough to show that $Q_T(b)$ is an upper semicontinuous function of b. Let $H : F(\infty) \to \mathbf{R}$ be defined by $H(g) = g_m(T)$. Then H is continuous and attains its supremum over $F(b) \cap A$ since the set $F(b) \cap A$ is compact by Lemmas 1 and 3.

Furthermore for any $y \in \mathbf{R}$, the set $S_y = \{g \in F(\infty) : H(g) < y\}$ is open. Then

$$\begin{aligned} \{b : Q_T(b) < y\} &= \{b : g_m(T) < y \text{ for all } g \in F(b) \cap A\} \\ &= \{b : F(b) \cap A \subset S_y\} \\ &= \{b : F(b) \subset S_y \cup A^c\}. \end{aligned}$$

So $\{b : Q_T(b) < y\}$ is an open set since $S_y \cup A^c$ is open and F is Berge-usc. ∎

The reader may wonder whether the sample throughput, TP_T for $T \in [0, \infty)$, is lower semicontinuous as well. In fact, TP_T is a discontinuous function of buffer capacities for finite T; see Figure 2. This is due to the fact that if two events occur at the same time, an infinitesimal change in buffer capacities may cause the order of these events to change, as illustrated by a simple, numerical example in Gürkan (1996a), p. 52-56. Of course, when the failure quantities and repair times for machines have continuous distributions, one may argue that the probability of a continuous random variable being equal to a specific value is zero; hence the probability that the time of two events coincides in a discrete event simulation is zero, as well. Therefore these types of phenomena cannot take place, in practice. On the other hand, it is clear from Figure 2 that once a sample path (a random number sequence ω) is fixed, there are some buffer capacities at which this type of phenomenon does occur and results in discontinuities in throughput. In other words, at each b the probability of throughput being discontinuous is zero; but the probability of throughput being discontinuous at *some* b is not zero.

Using the upper semicontinuity and monotonicity of sample throughput, one can easily show that for any finite T, any b, and any $\epsilon > 0$, there exists $\delta > 0$ such that for every j, if $0 < \Delta b_j < \delta$ and $b' = b + \Delta b_j$ then $TP_T(b) \leq TP_T(b') < TP_T(b) + \epsilon$. This shows that the phenomenon described in that example can occur when buffer capacities are *decreased* by an infinitesimal amount; it cannot occur when they are increased by an infinitesimal amount. Furthermore, this phenomenon may likewise occur when the operating time to failure (instead of operating quantity) is a random variable, see Remark 4.33 of Gürkan (1996a).

Fortunately, as can be seen in Theorem 8, the upper semicontinuity of TP_T suffices to prove the convergence of the conceptual method; the discontinuity of the sample functions does not constitute a problem from the theoretical point of view.

In the next result we use the term "proper" for an extended-real-valued function f. It means that f never takes the value $-\infty$ and it is not identically $+\infty$.

Remark Note that the analysis above does not depend on the particular distributions chosen for the random variables W_p^i and R_p^i. For the next result, Theorem 5, we assume that for each i and p, random variables W_p^i and R_p^i are exponentially distributed with means w_i and $1/r_i$ respectively, and show that $1/TP_T$ is a proper function of b. While proving this, we show that $TP_T(0) > 0$ for any T, which means

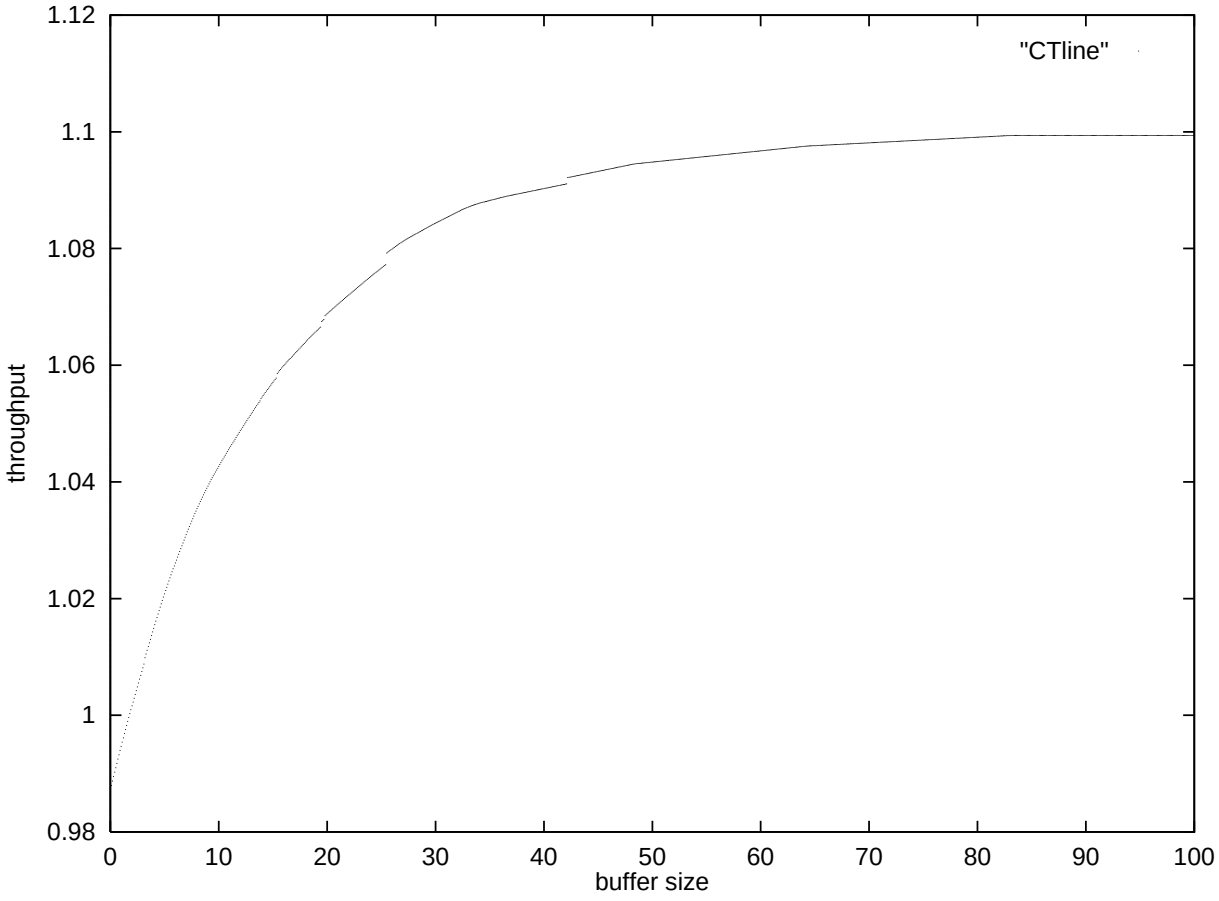

Figure 2: Simulation results for a 2-machine line with exponential failure and repair rates.

that a CT line with no buffer capacity has still positive throughput. Essentially any distribution whose support is on $(0,\infty)$ could be used in this result; the choice of the exponential distribution is made for ease of exposition.

Theorem 5 *For $T \in [0,\infty]$, $1/TP_T$ is a uniformly bounded, positive, proper function of b with probability one.*

Proof. First observe that for any T, TP_T is a bounded function of b; increasing the buffer capacities will not improve the line throughput beyond a certain point, since line throughput is bounded by C_m, the maximum flow rate of the last machine, in any case. We now show that for any T, $TP_T(0) > 0$. The properness will immediately follow since for any T and any b, $1/C_m \leq 1/TP_T(b) \leq 1/TP_T(0)$.

When $b = 0$, the line operates at the rate of the slowest machine, say C_{min} and it stops (i.e. fails) whenever one of the machines fails. Since there is no buffer between the machines and the product is continuous, this particular $m-$machine line degenerates to a $1-$machine line but with possibly more complicated failure and repair distributions. We have $TP_T(0) = Q_T/T$ where Q_T is the amount produced by this $1-$machine line in $[0,T]$. For this equivalent $1-$machine line, define

X_i: operating quantity between the $(i-1)st$ and ith failures of the machine,

Y_i: repair time after the ith failure.

Observe that the X_i are exponentially distributed random variables with rate $w_1^{-1} + \cdots + w_m^{-1}$ and the probability density function (pdf) of Y_i is given by

$$f(t) = \frac{w_1^{-1}}{w_1^{-1} + \cdots + w_m^{-1}} r_1 \cdot exp(-r_1 t) + \ldots + \frac{w_m^{-1}}{w_1^{-1} + \cdots + w_m^{-1}} r_m \cdot exp(-r_m t),$$

by conditioning on which machine has failed. Since $Q_T \geq \min\{X_1, C_{min}T\}$ and $X_1 > 0$ with probability one, we have $Q_T/T > 0$ with probability one for any finite T.

Let $t_0 = 0$ and $t_n =$ time of the nth repair for $n \geq 1$. Then $t_n = \sum_{i=1}^{n}(C_{min}^{-1}X_i + Y_i)$ and the amount produced at time t_n is $\sum_{i=1}^{n} X_i$. For any $T \in [t_{n-1}, t_n]$, the ratio Q_T/T is smallest either at $T = t_{n-1}$ or at $T = t_n$ (the quantity produced remains constant between $t_{n-1} + C_{min}^{-1}X_i$ and t_n). So

$$\inf_T \frac{Q_T}{T} = \inf_n \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n}(C_{min}^{-1}X_i + Y_i)}.$$

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to (w_1^{-1} + \cdots + w_m^{-1})^{-1} \text{ and } \frac{1}{n}\sum_{i=1}^{n} Y_i \to \frac{(r_1 w_1)^{-1} + \cdots + (r_m w_m)^{-1}}{w_1^{-1} + \cdots + w_m^{-1}} \text{ as } n \to \infty.$$

Hence

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n}(C_{min}^{-1}X_i + Y_i)} > 0.$$

So $K = \inf_T Q_T/T > 0$, and we conclude that $TP_T(0) \geq K > 0$, for any T. ∎

We now discuss how the assumptions of the conceptual method are satisfied under a regularity condition on the steady-state. First we give a general result about the epiconvergence of non-increasing functions, then apply it to the particular problem we have. The definition of epiconvergence (denoted by $\xrightarrow{e}$) can be found in Appendix A; the proof of Proposition 1 is provided in Appendix D.

Proposition 1 *Assume that with probability one,*
a. $L_n \to L_\infty$.
b. L_∞ *is lower semicontinuous.*
c. Each $L_n(1 \leq n \leq \infty)$ *is a non-increasing function.*
Then with probability one, $L_n \xrightarrow{e} L_\infty$.

Theorem 6 *Assume that with probability one,*
a. $TP_T \to TP_\infty$.
b. TP_∞ *is upper semicontinuous.*
Then with probability one, $1/TP_T \xrightarrow{e} 1/TP_\infty$.

Proof. Use Proposition 1 with Theorem 2. ∎

Theorem 6 shows that $1/TP_T \xrightarrow{e} 1/TP_\infty$, provided TP_∞ is upper semicontinuous. Intuitively, one even expects it to be continuous: the steady-state throughput of a line should not be very sensitive to small changes in the buffer capacities. In a 2-machine line, the continuity of steady-state throughput is provided by the analytical formula derived in Gershwin and Schick (1980). At this time we do not have a proof of the upper semicontinuity of TP_∞ for lines with more than 2 machines, although computational evidence strongly indicates that steady-state throughput is indeed a *continuous* function of buffer capacities. For an example, see Figure 3, which displays the throughput of a 2-machine CT line, where operating quantities to failures and repair times are exponentially distributed, for different run lengths T. In extensive numerical experiments (also for longer lines) we observed the same kind of behavior: a discontinuous function with frequent jumps of large sizes when T is small, but a smooth function when T is large.

As mentioned earlier, in an actual optimization problem, one would minimize a combination of the reciprocal of throughput and a cost function. The cost function usually captures the information regarding space limitations as well as costs of buffer capacities. In the next theorem, we show that the sample-path optimization method converges when applied to the optimization of throughput with respect to buffer capacities.

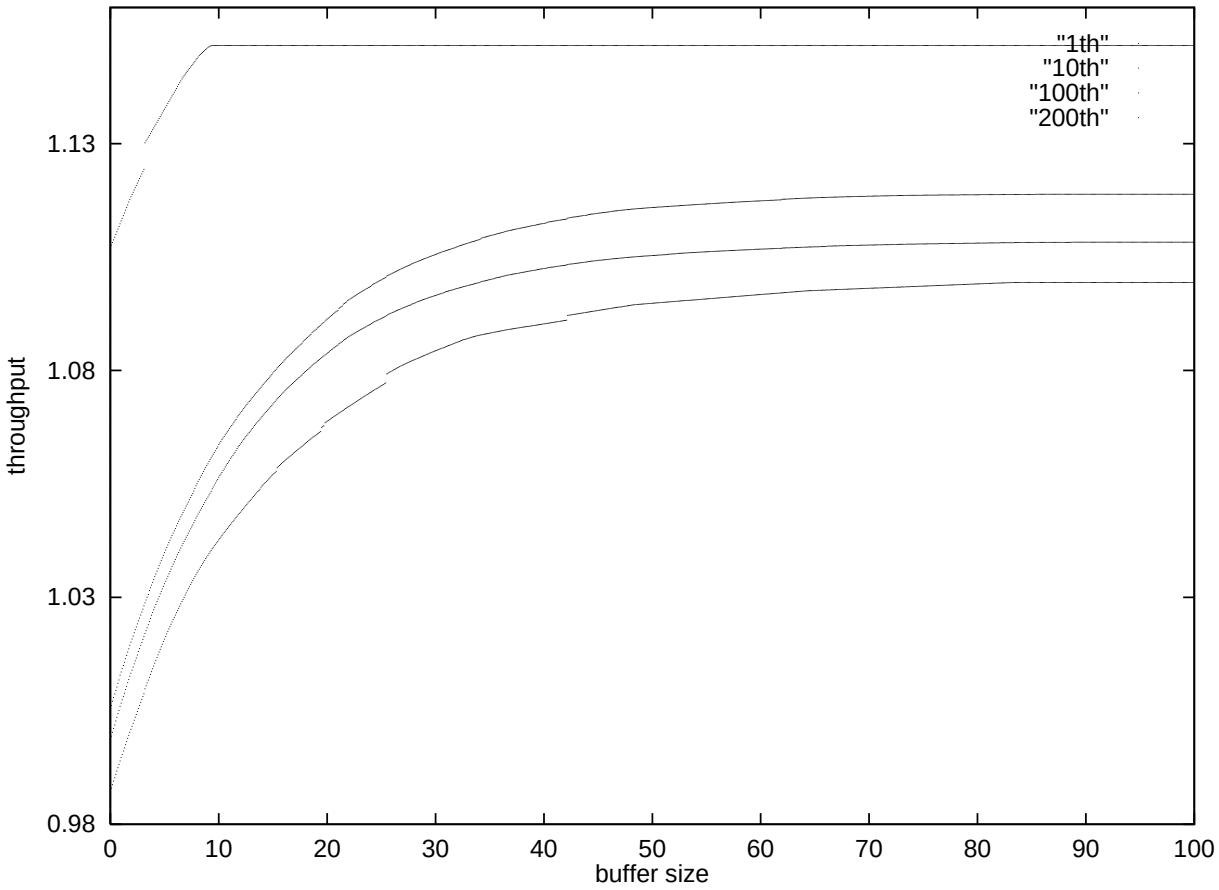

Figure 3: The throughput of a 2-machine CT line for different run lengths.

Theorem 7 *Suppose that TP_∞ is an upper semicontinuous function of b and f is a continuous, non-decreasing, non-negative function that is norm-coercive: i.e., $f(b) \to \infty$ as $||b|| \to \infty$. Let $\{f_T\}$ be a sequence of lower semicontinuous and proper cost functions associated with buffer capacities that converge uniformly on compact sets to f. Then for sufficiently large T and any positive scalars α and β, the set of minimizers of $\frac{\alpha}{TP_T} + \beta f_T$ is nonempty and any point in it is close to some minimizer of $\frac{\alpha}{TP_\infty} + \beta f$.*

Proof. The function $\frac{\alpha}{TP_\infty} + \beta f$ is lower semicontinuous, and it is norm-coercive because f is norm-coercive and $\frac{\alpha}{TP_\infty}$ is bounded below by zero. Thus the set of minimizers of $\frac{\alpha}{TP_\infty} + \beta f$ on $\mathbf{R}_+^{m-1}$ is nonempty and bounded; it must also be closed by lower semicontinuity. Use Theorems 4 and 5 to see that each $\frac{\alpha}{TP_T} + \beta f_T$ is a lower semicontinuous and proper function of b. Then apply Theorem 7.44 of Rockafellar and Wets (1996) and Theorem 6 to get $\frac{\alpha}{TP_T} + \beta f_T \xrightarrow{\text{e}} \frac{\alpha}{TP_\infty} + \beta f$. The result follows by Theorem 8 and Proposition 2. ∎

Remark Although Theorem 7 allows us to work with a sequence of functions $\{f_T\}$, a typical choice would be to use the constant sequence in which $f_T = f := \sum_{i=1}^{m-1} b_i$ for every T. This functional form would model a problem in which one wants to maximize the throughput but there are costs associated with increasing the buffer capacities.

4 Conclusion

In this paper we have discussed the theoretical issues that arise when applying a simulation-based method, sample-path optimization, to the buffer allocation problem in tandem lines with unreliable machines. We provided a novel mathematical framework to model the dynamics of the system and used this framework to prove the convergence of the conceptual method. As a by-product we established interesting properties of system throughput, such as monotonicity in buffer capacities and in machine flow rates. To the best of our knowledge, these structural properties are the first of their kind for the tandem lines studied in this paper.

The next question we shall address is a practical one: What do we actually do when we attempt to use the method in a real-world application? The second paper (Gürkan 1996b) will deal with the operational issues that arise in the implementation of this method. These include:

(1) How to use infinitesimal perturbation analysis to compute certain directional derivatives of sample throughput?

(2) What are the difficulties that arise during optimization due to the special structure of the sample functions (see Figure 2)?

(3) How does the sample-path optimization method perform numerically on this problem and what kind of modifications to the basic method can improve the

method's performance?

Acknowledgement

We thank our advisor S. M. Robinson and our friend Geoff Pritchard for many fruitful discussions on this subject.

Appendix A

This appendix contains the convergence results of Robinson (1996) that are relevant to the work presented here. We first need to define several crucial concepts.

Definition 1 *A sequence f_n of extended-real-valued functions defined on $\mathbf{R}^k$* epi-converges *to an extended-real-valued function f_∞ defined on $\mathbf{R}^k$ (written $f_n \xrightarrow{e} f_\infty$) if for each $\theta \in \mathbf{R}^k$ the following hold:*

a. For each sequence $\{\theta_n\}$ converging to θ, $f_\infty(\theta) \leq \liminf_{n\to\infty} f_n(\theta_n)$.

b. For some sequence $\{\theta_n\}$ converging to θ, $f_\infty(\theta) \geq \limsup_{n\to\infty} f_n(\theta_n)$.

Note that in (b) we actually have $f_\infty(\theta) = \lim_{n\to\infty} f_n(\theta_n)$, because of (a).

It is known that epiconvergence is independent of pointwise convergence in the sense that neither implies the other. For a very readable elementary treatment of the relationships between different types of convergence, see Kall (1986). Attouch (1984) contains comprehensive treatment of epiconvergence and related issues. Also see the forthcoming book Rockafellar and Wets (1996) for a treatment of epiconvergence from the perspective of optimization.

Definition 2 *Let Z be a topological space and let f be an extended-real-valued function on Z. A nonempty subset M of Z is a* complete local minimizing (CLM) set *for f with respect to an open set $G \supset M$, if the set of minimizers of f on $\mathrm{cl}\, G$ is M.*

The concept of a CLM set, introduced in Robinson (1987), extends the idea of an isolated local minimizer to cases in which the set of minimizers might not be a singleton.

Definition 3 *A multifunction F from a topological space Z to a topological space Y is* Berge-usc *at a point z_0 of Z if for each open set U of Y with $F(z_0) \subset U$ the set $\{z \in Z : F(z) \subset U\}$ is open. F is Berge-usc in Z if it is Berge-usc at every point of Z and if $F(z)$ is compact for every $z \in Z$.*

Berge-usc is introduced in Berge (1963) under the name "upper semicontinuity"; see Rockafellar and Wets (1996) for a treatment of relationships between various semicontinuity and continuity notions for multifunctions.

Let S and T be subsets of $\mathbf{R}^k$. We use the notation $e(S,T)$ for the *excess* of S over T, defined by

$$e(S,T) = \sup_{s\in S} d(s,T); \qquad d(s,T) = \inf_{t\in T} \|s-t\|.$$

If $e(S,T)$ small, then *each* point of S is close to *some* point of T, though some points of T might be far from any point of S. Such nonsymmetric behavior is not present in the *Hausdorff* distance between S and T (written $h(S,T)$) that is defined by $h(S,T) = \max\{e(S,T), e(T,S)\}$.

We can now state the basic result of Robinson (1996).

Theorem 8 [Theorem 3.7 of Robinson (1996)] *Suppose that the following assumptions hold:*

a. With probability one, each L_n $(1 \le n < \infty)$ is lower semicontinuous and proper.

b. With probability one, $L_n \xrightarrow{\text{e}} L_\infty$ as $n \to \infty$.

There is a subset Γ of Ω having measure zero, with the following properties: suppose that $\omega \notin \Gamma$, let G be an open bounded set in $\mathbf{R}^k$, define for $1 \le n \le \infty$

$$\hat{\mu}_n(\omega) = \inf_{\theta\in \mathrm{cl}\, G} L_n(\omega,\theta), \qquad \hat{M}_n(\omega) = \{\theta \in \mathrm{cl}\, G \mid L_n(\omega,\theta) = \hat{\mu}_n(\omega)\},$$

and assume that $\hat{M}_\infty(\omega)$ is a CLM set for $L_\infty(\omega,\cdot)$ with respect to G. Then

1. $\lim_{n\to\infty} \hat{\mu}_n(\omega) = \hat{\mu}_\infty(\omega)$, and $\hat{\mu}_\infty(\omega)$ is finite.

2. $\hat{M}_n(\omega)$ is Berge-usc at ∞, and $\hat{M}_\infty(\omega)$ is compact.

3. There is a finite positive integer N_ω such that for each $n \ge N_\omega$, $\hat{M}_n(\omega)$ is a nonempty, compact CLM set for $L_n(\omega,\cdot)$ with respect to G.

4. $\lim_{n\to\infty} e(\hat{M}_n(\omega), \hat{M}_\infty(\omega)) = 0$.

Theorem 8 permits us to look at sets of local minimizers that may not be global minimizers; in this sense its setting is very general. As explained in the next proposition, the assumption in Theorem 8 of the existence of a CLM set for $L_\infty(\omega,\cdot)$ can be replaced by a stronger, inf-compactness assumption.

Proposition 2 [Proposition 3.8 of Robinson (1996)] *Suppose that the following assumptions hold:*

a. With probability one, each L_n $(1 \le n < \infty)$ is lower semicontinuous and proper.

b. With probability one, $L_n \xrightarrow{\text{e}} L_\infty$ as $n \to \infty$.

c. With probability one, L_∞ is proper and its set M_∞ of minimizers is nonempty and compact.

Then for almost every ω, $M_\infty(\omega)$ is a CLM set for $L_\infty(\omega,\cdot)$ with respect to some open bounded set $G(\omega)$.

Some remarks are in order. First, in general the set G of Theorem 8 depends on the sample point ω, which may cause an inconvenience since we use this set to construct $\hat{\mu}_\infty(\omega)$ and $\hat{M}_\infty(\omega)$. This inconvenience can be removed by assuming that L_∞ is a deterministic function. This holds for the limit functions in this paper, since we consider steady-state throughput. Second, in the case of convex functions one can take G to be $\mathbf{R}^k$ in Theorem 8, i.e. the localization provided by G is not necessary. We refer the interested reader to Robinson (1996) for results in the case of convex functions. Finally, since numerical methods used in practice find solutions that are approximate, the behavior of the method when ϵ-minimizers are computed is quite important from a practical point of view. Results in Section 4 of Robinson (1996), especially Theorem 4.2, show that the behavior of the method remains unchanged in that case.

Appendix B

This appendix contains the proofs of the three technical lemmas from §3.1.

For $b \in \mathbf{R}_+^{m-1}$ write $F(b) = F_1(b) \cap F_2(b) \cap F_3(b) \cap F_4(b)$ where
$F_1(b) = \{g : g_1 \geq \ldots \geq g_m \geq 0\}$,
$F_2(b) = \{g : g_i \text{ is non-decreasing for each } i = 1, \ldots, m\}$,
$F_3(b) = \{g : g_i(0) = 0, |g_i(x) - g_i(y)| \leq C_i|x - y| \text{ for any } x, y \in [0, T] \text{ and } i = 1, \ldots, m\}$,
$F_4(b) = \{g : g_i(x) - g_{i+1}(x) \leq b_i \text{ for any } x \in [0, T], \text{ and } i = 1, \ldots, m-1\}$.

Lemma 1 *The multifunction F has the following properties:*
a. $\operatorname{gph} F$ *is closed.*
b. $\operatorname{gph} F$ *is convex.*
c. F is compact-valued and $F(b) \subset F(\infty)$ for all $b \in \mathbf{R}^{m-1}$.

Proof. For (a), we take a sequence $\{(b^n, g^n)\}$ in $\operatorname{gph} F$ that converges to a point (b, g) and show that $(b, g) \in \operatorname{gph} F$. Clearly, $g \in F_1(b) \cap F_2(b) \cap F_4(b)$. Take $\epsilon > 0$ and find a positive integer N_ϵ such that for all $n \geq N_\epsilon$, $t \in [0, T]$, and $i = 1, \ldots, m$, $\|g_i^n(t) - g_i(t)\| < \epsilon$. Then for all $x, y \in [0, T]$ and $i = 1, \ldots, m$,

$$\begin{aligned} \|g_i(x) - g_i(y)\| &= \|g_i(x) - g_i^n(x) + g_i^n(y) - g_i(y) + g_i^n(x) - g_i^n(y)\| \\ &\leq \|g_i(x) - g_i^n(x)\| + \|g_i(y) - g_i^n(y)\| + \|g_i^n(x) - g_i^n(y)\| \\ &< 2\epsilon + C_i\|x - y\|. \end{aligned}$$

Since ϵ can be made arbitrarily small, we must have $g \in F_3(b)$ as well. Hence $\operatorname{gph} F$ is closed.

To prove (b), we take $(b,g),(a,h) \in \operatorname{gph} F$ and $\lambda \in [0,1]$. Clearly,

$$(1-\lambda)g + \lambda h \in F_1((1-\lambda)b + \lambda a) \cap F_2((1-\lambda)b + \lambda a) \cap F_4((1-\lambda)b + \lambda a).$$

For any $i = 1, \ldots, m$ and $x, y \in [0,T]$,

$$\begin{aligned} \|[((1-\lambda)g_i(x) + \lambda h_i(x)] - [((1-\lambda)g_i(y) + \lambda h_i(y)]\| &\leq (1-\lambda)\|g_i(x) - g_i(y)\| + \lambda\|h_i(x) - h_i(y)\| \\ &\leq (1-\lambda)C_i\|x-y\| + \lambda C_i\|x-y\| \\ &= C_i\|x-y\|. \end{aligned}$$

Hence $(1-\lambda)g + \lambda h \in F_3((1-\lambda)b + \lambda a)$ as well.

Clearly, $F_1(b), F_2(b), F_3(b)$, and $F_4(b)$ are closed sets. Furthermore, for any $g \in F_3(b)$, any $x \in [0,T]$, and $i = 1, \ldots, m$, $|g_i(x)| \leq C_i|x| \leq C_iT$. Hence for any $g \in F_3(b)$, $\|g\| \leq \max_{i=1}^m C_iT$ and

$$\|g(x) - g(y)\| = \max_{i=1}^m |g_i(x) - g_i(y)| \leq \max_{i=1}^m C_i\|x-y\| \text{ for any } x, y \in [0,T].$$

Then by the Arzelà–Ascoli theorem $F_3(b)$ is compact. Hence F is compact-valued. Furthermore we have for all $b \in \mathbf{R}^{m-1}$, $F(b) \subset F(\infty)$. ∎

Lemma 2 *The multifunction F is Berge-usc in $\mathbf{R}^{m-1}$ and $b \mapsto F(b)$ is a continuous mapping from $\operatorname{int}(\mathbf{R}_+^{m-1})$ to compact subsets of $\mathcal{C}([0,T],\mathbf{R}^m)$ with the metric topology induced by the Hausdorff distance.*

Proof. Since $\operatorname{gph} F$ is closed and for all $b \in \mathbf{R}^{m-1}$, $F(b)$ is a subset of the compact set $F(\infty)$, the multifunction F is Berge-usc in $\mathbf{R}^{m-1}$ by the corollary to Theorem 7 in Section 7.1 of Berge (1963). Berge-usc implies that for any $\epsilon > 0$ and any $b \in \mathbf{R}^{m-1}$, there exists a $\delta > 0$ such that

$$e(F(b'), F(b)) < \epsilon \text{ for every } b' \text{ with } \|b' - b\| < \delta. \tag{4.1}$$

To see this, observe that $F(b) + \operatorname{int}(\epsilon\mathbf{B})$ is an open neighborhood of $F(b)$ and use the definition of Berge-usc.

Let $\epsilon > 0$ and take $b \in \operatorname{int}\operatorname{dom} F = \operatorname{int}(\mathbf{R}_+^{m-1})$ and $g \in F(b)$. By applying Theorem 1 of Robinson (1976) to the inverse multifunction F^{-1}, we can find $\delta(g) > 0$ such that $F^{-1}(g + \epsilon\mathbf{B}) \supset b + \delta(g)\epsilon\mathbf{B}$, i.e. if $\|b' - b\| < \epsilon\delta(g)$, then there exists $f \in F(b')$ with $\|g - f\| < \epsilon$. Notice that $\delta(g)$ depends on g; however for every $h \in F(b)$ one could always take $\delta(h) \geq \delta(g)(1 + \|h - g\|)^{-1}$, see p. 133 of Robinson (1976). If we let $K_g = \max_{h \in F(b)} \|h - g\|$ (which is attained since $F(b)$ is a compact set) and $\delta = \delta(g)(1 + K_g)^{-1}/\epsilon > 0$, then $\delta \leq \delta(h)$ for all $h \in F(b)$. So for all $g \in F(b)$ and b' with $\|b' - b\| < \delta$, there exists $f \in F(b')$ with $\|f - g\| < \epsilon$. This is equivalent to having $e(F(b), F(b')) < \epsilon$ if $\|b' - b\| < \delta$ which together with (4.1) gives the continuity of the mapping $b \mapsto F(b)$, for all $b \in \operatorname{int}(\mathbf{R}_+^{m-1})$ using the Hausdorff distance. ∎

Lemma 3 *A is closed in $F(\infty)$.*

Proof. Take a sequence $\{g^n\}$ in A that converges to a function g in $F(\infty)$. Assume that $g \notin A$. Then there exist i and j with $\lambda(\{t : g_i(t) = f_{ij}\}) < R^i_j$. Since $g \in F(\infty)$, each component of g is continuous and non-decreasing. Therefore the set $\{t : g_i(t) = f_{ij}\}$ is actually an interval, say $[r, s]$. Choose $\delta > 0$ small enough so that $\lambda([r-\delta, s+\delta]) < R^i_j$, g increases in $[r-\delta, r]$, and g increases in $[s, s+\delta]$. Then $\epsilon := \min\{g_i(s+\delta) - f_{ij}, f_{ij} - g_i(r-\delta)\} > 0$. Since the $g^n \to g$ in the sup-norm, we have uniform convergence in each component. Hence there exists N_ϵ such that $|g^n_i(t) - g_i(t)| < \epsilon$ for all $n \geq N_\epsilon$ and $t \in [0, T]$. Take $t > s + \delta$, then for any $n \geq N_\epsilon$ we have

$$\begin{aligned} g^n_i(t) &> g_i(t) - \epsilon \\ &\geq g_i(t) - (g_i(s+\delta) - f_{ij}) \\ &\geq f_{ij}. \end{aligned}$$

Similarly, we can show that $g^n_i(t) < f_{ij}$ for any $t < r - \delta$ and $n \geq N_\epsilon$. So for any $n \geq N_\epsilon$, if $t \notin [r-\delta, s+\delta]$ then $g^n_i(t) \neq f_{ij}$. Therefore we have

$$\lambda(\{t : g^n_i(t) = f_{ij}\}) \leq \lambda([r-\delta, s+\delta]) < R^i_j,$$

by choice of δ. This contradicts the fact that $g^n \in A$. ∎

Appendix C

This appendix contains the proof of Theorem 1 from §3.1.

Theorem 1 *Suppose that the event times have no cluster point. Then for each finite time T, $Q_T(b) = q_m(T)$.*

Proof. Let $v_i(t)$ be the rate of machine i at time t under strategy q and $v^g_i(t)$ be the rate of machine i at time t under strategy g. When t is the time of an event, we take $v_i(t) = v_i(t^+)$.

Without loss of generality we assume $b_i > 0$ for each i (otherwise we could combine two machines). Suppose there exists $g \in F(b) \cap A$ such that $g_m(T) > q_m(T)$. Let $\tau = \inf\{t : g_i(t) > q_i(t) \text{ for some } i\}$ where $\tau < T$. Suppose that $\{t_k\}$ is a sequence decreasing to τ, such that for each k there is an index i_k with $g_{i_k}(t_k) > q_{i_k}(t_k)$. By using the pigeonhole principle we can find some i such that for a subsequence $\{t_{k_j}\}$ we have $g_i(t_{k_j}) > q_i(t_{k_j})$ for each j. For simplicity, rename this sequence as $\{t_k\}$. Note that $g_i(\tau) = q_i(\tau)$ and $g_i(t) > q_i(t)$ for $t \in (\tau, \tau + \delta_0]$ for some $\delta_0 > 0$, by continuity of g_i and q_i.

Under strategy q, machine i cannot be under repair at time τ. To see this, suppose it were not true; then under strategy g machine i must have finished the same repair by time τ. So it must have begun the repair earlier, say at t_0, whereas under q machine i began its repair at time $t_1 > t_0$. But $q_i(t) < q_i(t_1)$ for $t < t_1$ (failures are operational only), so $g_i(t_0) = q_i(t_1) > q_i(t_0)$ which contradicts the definition of τ.

By assumption, τ is not a cluster point of the event times. Since under q the rate of machine i changes only at an event time, there is $\delta_1 > 0$ such that in the interval $[\tau, \tau + \delta_1]$ that rate is constant, say v_i^q. We claim that $v_i^q < C_i$. To see this, observe that if it were not true, then we would have for all $\delta \in (0, \min\{\delta_0, \delta_1\})$

$$g_i(\tau + \delta) - q_i(\tau + \delta) = g_i(\tau) - q_i(\tau) + \int_\tau^{\tau+\delta} [v_i^g(t) - C_i]dt.$$

Since $g_i(\tau) = q_i(\tau)$, we would have $g_i(\tau + \delta) - q_i(\tau + \delta) \leq 0$ which contradicts the existence of δ_0.

Therefore for small enough $\delta \in (0, \min\{\delta_0, \delta_1\})$ either

a) $q_i(t) = q_{i-1}(t)$ for $t \in [\tau, \tau + \delta]$

or

b) $q_i(t) = q_{i+1}(t) + b_i$ for $t \in [\tau, \tau + \delta]$;

since if neither (a) nor (b) occurs, then machine i should be running at rate C_i on $[\tau, \tau + \delta)$.

If (a) occurs, then for sufficiently large k

$$\begin{aligned} g_{i-1}(t_k) - q_{i-1}(t_k) &= g_{i-1}(t_k) - g_i(t_k) - [q_{i-1}(t_k) - q_i(t_k)] + g_i(t_k) - q_i(t_k) \\ &\geq g_i(t_k) - q_i(t_k) > 0. \end{aligned}$$

We get the first of these inequalities since $g_{i-1}(t_k) - g_i(t_k) \geq 0$ and $q_{i-1}(t_k) = q_i(t_k)$. The second inequality is a consequence of the choice of δ. Now we can repeat the same argument for machine $i - 1$. Note that we must then have $g_{i-1}(\tau) = q_{i-1}(\tau)$ and this time we know that only (a) can occur. So we get the same property for $i - 2,\ i - 3, \ldots$. Eventually we reach machine 1 and a contradiction (since the first machine is never starved).

If (b) occurs, then for sufficiently large k

$$\begin{aligned} g_{i+1}(t_k) - q_{i+1}(t_k) &= g_{i+1}(t_k) - g_i(t_k) + q_i(t_k) - q_{i+1}(t_k) + g_i(t_k) - q_i(t_k) \\ &\geq g_i(t_k) - q_i(t_k) > 0. \end{aligned}$$

The first of these inequalities follows from $g_{i+1}(t_k) + b_i \geq g_i(t_k)$ and $q_i(t_k) - q_{i+1}(t_k) = b_i$. The second inequality is a consequence of the choice of δ. Here again we must have $g_{i+1}(\tau) = q_{i+1}(\tau)$. Therefore we can repeat the above argument for machine $i + 1$ and this time we know that (b) is the only possibility. So we get the same property for $i + 2,\ i + 3, \ldots$. Eventually we reach machine m and a contradiction (since the last machine is never blocked). ∎

Appendix D

This appendix contains the proof of Proposition 1 from §3.2.

Proposition 1 *Assume that with probability one,*
a. $L_n \to L_\infty$.
b. L_∞ *is lower semicontinuous.*
c. Each $L_n (1 \le n \le \infty)$ *is a non-increasing function.*
Then with probability one, $L_n \xrightarrow{e} L_\infty$.

Proof. Construct a set Γ of measure zero such that whenever $\omega \notin \Gamma$, $L_n \to L_\infty$ pointwise, L_∞ is lower semicontinuous, and for each $n = 1, \ldots, \infty$, L_n is a non-increasing function. Choose any $\omega \notin \Gamma$ and for brevity omit the sample point ω from L_n and L_∞.

We first prove that L_n are (almost) equi-lower semicontinuous, i.e for any $x \in \mathbf{R}^n$ and $\epsilon > 0$ there exist a neighborhood $U(x, \epsilon)$ and a number $N(x, \epsilon)$ such that

$$L_n(y) > L_n(x) - \epsilon \text{ for each } y \in U(x, \epsilon) \text{ and } n \ge N(x, \epsilon).$$

Fix x and $\epsilon > 0$. Since L_∞ is lower semicontinuous, we can find a $\delta > 0$ satisfying $L_\infty(y) > L_\infty(x) - \epsilon/3$ for $y \in \Pi_{i=1}^m [x_i - \delta, x_i + \delta]$. We also have $L_n(x + \delta) \to L_\infty(x + \delta)$ and $L_n(x) \to L_\infty(x)$ by pointwise convergence where $(x + \delta)$ means $(x_1 + \delta, \ldots, x_m + \delta)$. Hence we can choose N such that

$$L_n(x + \delta) > L_\infty(x + \delta) - \epsilon/3 \text{ and } L_\infty(x) > L_n(x) - \epsilon/3 \text{ for } n \ge N.$$

Then for $n \ge N$,

$$L_n(y) \ge L_n(x+\delta) > L_\infty(x+\delta) - \epsilon/3 > L_\infty(x) - \epsilon/3 - \epsilon/3 \ge L_n(x) - \epsilon/3 - 2\epsilon/3 = L_n(x) - \epsilon.$$

Now $L_n \xrightarrow{e} L_\infty$ follows from Theorem 5 of Kall (1986). ∎

References

[1] Anantharam, V. and Tscoucas, P. 1990. Stochastic concavity of the throughput in series of queues with finite buffers. *Advances in Applied Probability* 22: 761-763.

[2] Attouch, H. 1984. *Variational Convergence for Functions and Operators* (Pitman, Boston).

[3] Berge, C. 1963. *Topological Spaces* (The MacMillan Company, New York).

[4] Buzacott, J.A. and Shanthikumar, J.G. 1992. Design of manufacturing systems using queueing models. *Queueing Systems* 12: 135-214.

[5] Caramanis, M. 1987. Production system design: A discrete event dynamic system and generalized Benders' decomposition approach. *International Journal of Production Research* 25(8): 1223-1234.

[6] Correa, R. and Lemaréchal, C. 1993. Convergence of some algorithms for convex minimization. *Mathematical Programming* 62: 261-275.

[7] Fletcher, R. 1987. *Practical Methods of Optimization, 2nd Ed.* (Wiley, Chichester).

[8] Fu, B.-R. 1996. Modeling and analysis of discrete tandem production lines using continuous flow models. Ph. D. Dissertation. Department of Industrial Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA.

[9] Fu, M.C. and Healy K. 1992 Simulation optimization of (s, S) inventory systems. In: *Proceedings of the 1992 Winter Simulation Conference*, eds. J.J. Swain, D. Goldsman, R.C. Crain and J.R. Wilson (IEEE, Piscataway, New Jersey), 506-514.

[10] Gershwin, S.B. 1987. An efficient decomposition method for the approximate evaluation of production lines with finite storage spaces. *Operations Research* 35: 291-305.

[11] Gershwin, S.B. and Schick, I.C. 1980. Continuous model of an unreliable two-stage material flow system with a finite buffer. Technical Report LIDS-R-1039. Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[12] Gershwin, S.B. and Schick, I.C. 1983. Modeling and analysis of three-stage transfer lines with unreliable machines and finite buffers. *Operations Research* 31(2): 354-380.

[13] Gill, P.E., Murray, W., and Wright, M.H. 1981. *Practical Optimization* (Academic Press, London).

[14] Glasserman, P., and Yao, D.D. 1992a. Monotonicity in generalized semi-Markov processes. *Mathematics of Operations Research* 17(1): 1-21.

[15] Glasserman, P., and Yao, D.D. 1992b. Generalized semi-Markov processes: Antimatroid structure and second order properties. *Mathematics of Operations Research* 17(2): 444-469.

[16] Gürkan, G. 1996a. Performance optimization in simulation: Sample-path optimization of buffer allocations in tandem lines. Ph. D. Dissertation. Department of Industrial Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA.

[17] Gürkan, G. 1996b. Sample-path optimization of buffer allocations in a tandem queue - Part II: Operational issues and implementation. Working Paper. CentER for Economic Research, Tilburg University, Tilburg, The Netherlands.

[18] Gürkan, G., Özge, A.Y. and Robinson, S.M. 1994. Sample-path optimization in simulation. In: *Proceedings of the 1994 Winter Simulation Conference*, eds. J.D. Tew, S. Manivannan, D.A. Sadowski and A.F. Seila (IEEE, Piscataway, New Jersey), 247-254.

[19] Gürkan, G., Özge, A.Y. and Robinson, S.M. 1996. Sample-path solution of stochastic variational inequalities, with applications to option pricing. To appear in: *Proceedings of the 1996 Winter Simulation Conference*.

[20] Hillier, F.S. and So, K.C. 1991. The effect of machine breakdowns and interstage storage on the performance of production line systems. *International Journal of Production Research* 29(10): 2043-2055

[21] Ho, Y.-C., Eyler, M.A. and Chien, T.T. 1979. A gradient technique for general buffer storage design in a production line. *International Journal of Production Research* 17(6): 557-580.

[22] Ho, Y.-C., Eyler, M.A. and Chien, T.T. 1983. A new approach to determine parameter sensitivities of transfer lines. *Management Science* 29(6): 700-714.

[23] Kall, P. 1986. Approximation to optimization problems: An elementary review. *Mathematics of Operations Research* 11: 9-18.

[24] Kiefer, J. and Wolfowitz, J. 1952. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics* 23: 462-466.

[25] Kiwiel, K.C. 1990. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming* 46: 105-122.

[26] L'Ecuyer, P., Giroux, N. and Glynn, P.W. 1994. Stochastic optimization by simulation: numerical experiments with the M/M/1 queue in the steady-state. *Management Science* 40: 1245-1261.

[27] Loynes, R. 1962. The stability of a queue with non-independent inter-arrival and service times. *Proccedings of the Cambridge Philosophical Society* 58: 497-520.

[28] Meester, L.E. and Shanthikumar, J.G. 1990. Concavity of the throughput of tandem queueing systems with finite buffer space. *Advances in Applied Probability* 22: 764-767.

[29] Meketon, M. 1983. A tutorial on optimization in simulations. Unpublished tutorial presented at the 1983 Winter Simulation Conference.

[30] Meketon, M. 1987. Optimization in simulation: A survey of recent results. In: *Proceedings of the 1987 Winter Simulation Conference*, eds. A. Thesen, H. Grant, and W.D. Kelton (IEEE, Piscataway, New Jersey), 58-67.

[31] Moré, J. and Wright, S.J. 1993. *Optimization Software Guide.* Frontiers in Applied Mathematics Vol.14 (Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania).

[32] Nummelin, E. 1981. Regeneration in tandem queues. *Advances in Applied Probability* 13: 221-230.

[33] Plambeck, E.L., Fu, B.-R., Robinson, S.M., and Suri, R. 1996. Sample-path optimization of convex stochastic functions. Accepted by *Mathematical Programming.*

[34] Robbins, H. and Monro, S. 1951. A stochastic approximation method. *Annals of Mathematical Statistics* 22: 400-407.

[35] Robinson, S.M. 1976. Regularity and stability for convex multivalued functions. *Mathematics of Operations Research* 1(2): 130-143.

[36] Robinson, S.M. 1987. Local epi-continuity and local optimization. *Mathematical Programming* 37: 208-222.

[37] Robinson, S.M. 1996. Analysis of sample-path optimization. *Mathematics of Operations Research* 21(3): 513-528.

[38] Rockafellar, R. T. and Wets, R. J-B. 1996. *Variational Analysis* (Springer-Verlag Berlin), forthcoming.

[39] Schassberger, R. 1976. On the equilibrium distribution of a class of finite-state generalized semi-Markov processes. *Mathematics of Operations Research* 1: 395-406.

[40] Schramm, H., and Zowe, J. 1990. A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results. Technical Report, 209. Mathematisches Institut, Universität Bayreuth, Bayreuth, Germany.

[41] Shanthikumar, J.G. and Yao, D.D. 1989a. Stochastic monotonicity in general queueing networks. *Journal of Applied Probability* 26: 413-417.

[42] Shanthikumar, J.G. and Yao, D.D. 1989b. Second-order stochastic properties in queueing systems. *Proceedings of IEEE* 77(1): 162-170.

[43] Suri, R. and Fu, B.-R. 1994. On using continuous flow lines to model discrete production lines. *Discrete Event Dynamic Systems* 4: 129-169.

[44] Yamashita, H. and Önvural, R.O. 1994. Allocation of buffer capacities in queueing networks with arbitrary topologies. *Annals of Operations Research* 48: 313-332.